Center for
Educator Compensation Reform

DATA QUALITY ESSENTIALS

Jeffery G. Watson, University of Wisconsin-Madison

Editor, Cynthia D. Prince, Vanderbilt University.

The Center for Educator Compensation Reform (CECR) was awarded to Westat—in partnership with Learning Point Associates, Synergy Enterprises Inc., Vanderbilt University, and the University of Wisconsin—by the U.S. Department of Education in October 2006.

The primary purpose of CECR is to support the Teacher Incentive Fund grantees with their implementation efforts through the provision of ongoing technical assistance and the development and dissemination of timely resources. CECR also is charged with raising national awareness of alternative and effective strategies for educator compensation through a newsletter, a Web-based clearinghouse, and other outreach activities.

This work was originally produced in whole or in part by the CECR with funds from the U.S. Department of Education under contract number ED-06-CO-0110. The content does not necessarily reflect the position or policy of CECR or the Department of Education, nor does mention or visual representation of trade names, commercial products, or organizations imply endorsement by CECR or the federal government.

DATA QUALITY ESSENTIALS

In order to complete compensation reform successfully, many school systems must transform information systems that were originally designed for reporting and accountability into systems that support performance-pay work. However, using data systems in new ways can quickly expose previously unnoticed data quality problems. The goal of this article is to help school systems identify, address, and plan for data quality problems before performance decisions are put under the scrutiny of system stakeholders.

This module focuses on the data quality challenges that states, districts, and schools must resolve when they reform compensation systems to take into account performance measures such as student achievement, teacher evaluation, and professional development. To begin, the following key questions must be addressed:

1. What are the key characteristics of quality data for compensation reform projects?
2. What are some common ways in which data quality problems can manifest within a compensation reform project?
3. What are some potential solutions to those problems?

The Teacher Incentive Fund (TIF) has awarded more than $80 million to 34 local and state education agencies to support the design and implementation of performance-pay systems. In addition to TIF grantees, many other school systems are also examining and implementing performance-pay systems. These projects all have heavy information technology (IT) components because districts generally use measures drawn from many data sources to determine individual pay amounts, including assessments, student enrollment, human resources, and teacher and principal observations. Because districts use these extensive systems to make a relatively small number of decisions, there is the inherent tendency for what would be otherwise isolated data quality issues to be magnified within the performance plans.

For example, most districts that base teacher pay simply on years of experience and highest degree earned only have to worry about using their information systems within a fairly constrained scope. These districts generally use their student information systems only to enroll students in schools, schedule students and teachers into course sections, and track attendance and possibly disciplinary actions. They generally use their human resource systems only to track employee data and deliver payroll, and their assessment systems only to meet accountability requirements at the school and grade level. However, school systems implementing compensation reform must use all of these systems to make a single decision: how much to pay teachers and principals. Data quality problems from any one system have the potential to affect a compensation reform project in negative ways.

Data quality problems from any one system have the potential to affect a compensation reform project in negative ways.

The goals of this article are to identify the dimensions of data quality from a compensation reform perspective, identify common data quality problems, and offer solutions to those problems. To address the first goal, this article presents six dimensions of data quality that are key to understanding how information systems must evolve to meet the needs of education leaders who are developing new pay systems.[i] Individually, these dimensions represent design requirements for overhauling district data systems for the purpose of building decision support capacity. Based on work with several large U.S. districts across multiple projects, these dimensions focus attention on the functional role of data and information systems within decision making. These dimensions do not specify a data model per se, nor do they specify content (e.g., prescribe a data dictionary). The dimensions complement the work by the Schools Interoperability Framework Association (SIFA) and the Data Quality Campaign by focusing attention on the role of data within the context of decision making (i.e., determining performance awards) and the technology environments of large districts.[ii]

What characteristics does a data system need to have in order to support a performance-based pay system?

Watson has identified six dimensions of data quality: accuracy, validity, granularity, interoperability, relational, and reducibility.[1] The definitions of these dimensions are presented below.

1. **Accuracy** is the degree to which data reflect reality. Are the data correct? This is a fundamental aspect of data quality and is probably one that easily comes to mind for most people when they confront the issue of data quality. An example of poor accuracy comes from a large urban district that recently attempted to merge teacher certification data from its human resource system with teacher course assignment data from its student information system. In theory, records for the teachers in these two systems should

[i] These six data quality dimensions were first presented by Jeffery Watson, University of Wisconsin-Madison, at the National and International Workshop on School Information Systems and Data Based Decision Making. A copy of his conference manuscript, titled *Defining Data Quality for Decision Support Systems in Education*, can be found in the proceedings of that conference.

[ii] The Schools Interoperability Framework (SIF) is a data-sharing specification originally developed to allow information systems within K-12 districts to exchange data without requiring wholesale replacement of existing systems. It includes both clear definitions for core data elements, as well as secure methods for exchanging data. The Schools Interoperability Framework Association was founded to define the original standard and to provide a governance infrastructure for improving and expanding the standards-setting work. The SIF Association includes private software firms, state educational agencies, school districts, and higher education institutions. The association has also expanded to include international members. See http://www.sifinfo.org for more information on the standard, the association, and its members.

have matched, but in reality, only about 80 percent of the records matched on name or identification number. Common causes of inaccurate data include poorly designed computer interfaces, inadequate training, and human error.[2]

2. **Validity** is the degree to which data measure an intended construct. Do the data, regardless of their accuracy, represent the attribute or variable that they are supposed to represent? In a simple example, the U.S. Census Bureau changed the way in which race and ethnicity are reported. Instead of using one variable to report both race and ethnicity (e.g., White, Black, Hispanic), the Bureau now reports race and ethnicity separately, so that it is now possible to differentiate race and ethnicity independently (e.g., Black Hispanic versus White Hispanic).

   One common example in education is the school of record for any given student. While most students do not change schools during an academic year, many do, especially in urban settings. Thus, the school at which students are tested may not be the school at which they received most of their instruction. Because school-level student achievement measures become increasingly invalid as the number of mobile students increases, many districts will hold schools accountable only for those students who were enrolled for a full academic year. In this case, student achievement measures for a given school lose validity as the percentage of mobile students increases.

Six Dimensions of Data Quality	
Accuracy	The degree to which data reflect reality.
Validity	The degree to which data measure an intended construct.
Granularity	The number of individuals (e.g., students), items (e.g., test questions), or period of time (e.g., semester versus yearly attendance) over which data are aggregated.
Interoperability	The degree to which data are integrated across data systems.
Relational	The degree to which an information system's underlying model of a data system is capable of capturing reality.
Reducibility	The degree to which data support the formation of categories of entities.

3. **Granularity** is the number of individuals (e.g., students), items (e.g., test questions), or period of time (e.g., semester versus yearly attendance) over which data are aggregated. Data quality suffers when the granularity of data does not support the analytic lens, or unit of analysis, of decision makers.[3] For example, in urban districts, student mobility is often cited as a problem for schools because students who are mobile are exposed to disjointed instruction and curricula. Attempting to control for the amount of time a mobile

student spends between two schools requires student-school data to be sampled frequently. However, many districts capture student-school linkages between one and three times per year (usually for determining budgets). Under-sampling student-school linkages only limits the degree to which student learning can be attributed to schools or teachers.

4. **Interoperability** is the degree to which data are integrated across data systems. Generally, information systems in school districts are not integrated, although the SIFA has made significant progress toward establishing a unifying data model for developers. However, most school systems do not currently have a high degree of interoperability between source systems.

   There are many reasons why data quality usually suffers when systems are not integrated. First, when systems are not interoperable, data migration is cumbersome. For example, if a student information system is not integrated with the human resource system, staff must enter teacher data twice, which increases the likelihood of spelling and typographical errors. In addition, if teachers change their last names when they get married, staff must update teacher data in two systems, rather than just one. This would require staff to match records between the two systems using a combination of automated and manual methods, a process that is likely to be both expensive and difficult.

5. **Relational** is defined as the degree to which an information system's underlying model of a data system is capable of capturing reality. When a data model is not able to capture the state of affairs within a school, there is little hope that the data system will provide data that reflect what really happened within that school. For example, many student information systems do not capture alternative approaches to course scheduling. Most systems allow schools to enter one teacher assignment for each course. When teachers decide to team-teach, or otherwise collaborate during instruction, it becomes difficult, if not impossible, to record teacher assignments accurately. Other examples of scheduling approaches that are difficult to capture from student information systems include block scheduling, remediation interventions (e.g., pull-out instruction, tutoring), and special education instruction.

6. **Reducibility** is the degree to which data support the formation of categories of entities. For example, teachers are often labeled as math or science teachers, or as a teacher of a particular grade. Categorizing teachers as either math or science teachers when they actually teach across content areas would be an over-reduction of teacher assignment data. Likewise, assigning one school code to

students who are mobile is an over-reduction of student enrollment data. Many times the causes of over-reduction of data lie in how data are pulled from source systems and pushed into a repository (e.g., a data warehouse). That is to say, the over-reduction of data (e.g., excluding mobile students' alternate schools) sometimes occurs downstream of the student information system.

These dimensions should be used to foster dialogue between program directors, district policy makers, and IT staff. Ideally, administrators will engage IT staff in early discussions about these data quality dimensions for all sources of data that will be used to determine performance awards. Only staff who are intimately familiar with the systems in play will be able to assess many of these aspects of data quality accurately. Without this kind of collaboration, projects risk incorrectly awarding bonuses. The result of such a misstep could be catastrophic.

Watson proposes several types of questions to consider when applying the data quality dimensions to performance-based pay systems. Table 1 presents these questions to provide school systems with a sense of how they might begin to have conversations with IT staff.

Table 1. Using data quality dimensions to guide discussions between project leaders and information technology staff.

Data quality dimensions	Sample questions to ask information technology staff
Accuracy	Are student-teacher linkages in the student information system (SIS) correct? Do teacher records in SIS match teacher records in the human resources (HR) system?
Granularity	Do data support using a unit of analysis that matches the performance-pay systems (e.g., individual teacher bonuses)?
Validity	Are performance metrics consistent with other performance measures? Do student-teacher links captured in SIS reflect those in classrooms?
Interoperability	Can students be connected to teachers and other instructional staff? How much work will be involved in making sure that individuals (e.g., students and teachers) match across systems?
Relational	Is the SIS data model able to capture secondary student-teacher linkages?
Reducibility	Are teachers of multiple subjects incorrectly identified as teachers of only one particular content area? Do categories represent all teachers?

Assessing and Improving Data Quality

When school district staff encounter data quality problems, they may be tempted to ignore them, but to do so risks losing stakeholder support for the project. If data quality problems arise after awards are paid out, reactions will likely be very negative. We strongly caution school systems implementing compensation reform to anticipate and plan for data quality problems that may arise. Solutions are usually within reach, though project staff and IT staff will need to support corrective actions jointly.

Most likely, data quality problems will have both social and technological roots.[4] For example, it might be tempting to blame inaccurate data on sloppy data processing, but it is important to assess whether the interface design of a data system causes an increase in data entry errors. More training may be needed for technology staff, or work pressures may encourage users to take shortcuts or otherwise subvert the system from its intended use. Regardless of the causes, school systems should consider prioritizing data quality issues to help guide efforts to improve data quality. Usually long- and short-term solutions will be identified.

School systems implementing compensation reform must anticipate and plan for data quality problems that may arise. Solutions are usually within reach.

The remainder of this module presents three examples from actual school districts to illustrate how these dimensions can be used to identify, assess, and improve data quality.

Data Quality Challenge #1

Connecting teacher data from separate student information and human resource systems

Schools will have to merge data from the student information systems with data from their payroll systems. As noted previously, integrating data across these systems is not always easy. In one large urban district, only about 80 percent of the teacher records in the district's human resources system matched teacher records in the student information system. Some of the actual teacher 'names' that were entered in the student information system that could not be linked to actual teachers included:

Teacher A – MRP2

Teacher B – MRP1

Teacher C – Sci6B

Teacher D – Orchestra

In addition, some buildings used organizational structures that were not manageable with the student information system because the underlying data model did not allow teachers to assume multiple roles.

Two types of analyses should be helpful for determining the extent to which inaccurate data compromise a district's ability to integrate data across systems. First, matching teachers in the student information system and the human resource system reveals when data accuracy is lacking, because inaccurate data will result in incomplete matches. It may be helpful to summarize matches by grade and school. Second, understanding why inaccurate data are occurring involves analyzing workflows that might affect data quality. For example, in prior work with a large urban district, the author identified multiple factors that led to poor data quality.

One glaring cause was that the two systems used two different identification systems, which required data processing staff to look up teachers' employee numbers and names in the human resource system and enter these data by hand into the student information system. If these systems were better integrated, teacher identifiers would in most cases load into the student information system directly from the human resource system. Another factor was found when data entry duties were analyzed. As is the case in most districts, schools followed a complex workflow that required data processing and administrative staff to create course catalogues and preliminary schedules well before it was known who would actually teach at each school. Thus, staff sometimes had to create placeholders for teachers who would be hired in the future. Staff had to update teacher assignments once staffing was finalized (sometimes after the beginning of the school year). Failure to update the student information system correctly resulted in entries like 'Teacher A – MRP2.'

One of the most pressing concerns for IT system builders in light of the requirements of No Child Left Behind (NCLB) *legislation is creating the capacity to demonstrate improvement in student learning outcomes using longitudinal test data.*

Potential Solutions to Data Quality Challenge #1

There are four potential solutions to these problems. They involve both social and technical interventions and both short- and long-term interventions.

1. Build data quality checks for data-entry screens that use heuristics or look-up tables. For example, when an employee number is entered into a system, the system could check the number to make sure it conforms to the expected format (e.g., the correct number of digits). Better yet, data entry should be minimized whenever possible by pulling data from other systems rather than requiring the same information to be re-entered into a secondary system.
2. Create data quality management tools (e.g., reports, training procedures) for district administrators to identify schools that need to improve data quality.
3. Build support for data entry staff (e.g., training, tech support).
4. Identify true needs of schools and develop use-cases in order to provide feedback to the student information system vendor and improve the underlying student information system data model (e.g., scheduling logistics).

Data Quality Challenge #2

Connecting teachers to students

Student information systems often do not support the complex organizational structures that schools use. Schools use a variety of organizational designs.

Knowing which teachers taught which students is a critical linkage for school systems. However, student information systems often do not support the complex organizational structures that schools use. Schools use a variety of organizational designs, such as team-teaching and pullouts, but the data model of student information systems may not capture these non-traditional instructional models. Moreover, the data model may not capture additional instructional support staff (e.g., pull-out specialists, instructors in after-school activities). Districts often do not worry about capturing all of the nuances of a student's school year. Instead, they focus only on identifying a teacher (and school) of record and ignore other instructors who also contribute to student learning.

In addition, student information systems often do not record multiple roles of individuals or flexible organizational units. Ideally, a student information system should:

1. Link mobile students to multiple teachers who contributed to their learning gains;
2. Use course titles that reflect true curricular content;
3. Indicate when team-teaching is occurring and who teaches what; and
4. Link students to additional staff who provide instruction (for example, in pull-outs, tutoring, and after-school programs), not just classroom teachers of record.

Assessing the accuracy and validity of the student-teacher linkages is a good first step toward knowing the extent to which this particular data quality issue presents challenges to a performance-pay system. One way to do this is as follows. First, identify where student-teacher linkages are easiest to track (e.g., elementary schools that use traditional organizational models). This simplifies the problem and makes analysis more manageable. Second, count the number of students assigned to each teacher and identify any outliers, such as teachers with too many or too few students. Third, examine these outliers more closely for patterns (e.g., some special education teachers may have taught a small number of students across multiple sites) and ask administrators to verify or correct the information.

Potential Solutions to Data Quality Challenge #2

Three solutions to student-teacher linking problems were identified:

1. Build management tools, such as reports, that summarize student-teacher linkages (e.g., student counts by teacher, counts of teachers per building, identify which teachers teach across schools) and that can be used to target training and management solutions.

2. Examine the capacity of the student information system to track students' exposure to team-teaching, block scheduling, and pull-outs. Consider alternatives for collecting data from schools when these strategies are used. This solution helps assess the validity of the data.

3. Create incentives for schools to record the teacher of record accurately and verify this with teachers. For example, a district might require teachers to build a course roster from a list of enrolled students. Although this is redundant, it serves to validate the accuracy of the teacher-student links in the district's student information system. This solution also helps improve the validity of the data. Guilford County, North Carolina, for example, piloted a student-teacher linkage verification process in 2007 to provide opportunities for all teachers to review and confirm the names of each student that they taught. Although most student-teacher linkages were correct, teachers did catch some errors before the district performed the final analyses that would be used to determine performance awards. The process included three rounds of data verification to ensure that the linkages were accurate.[5]

Data Quality Challenge #3

Classifying teachers into categories

Middle school enrollment data from a large urban district provide an excellent example for this type of data quality challenge. The extent to which teachers instruct across grades and content areas can be assessed by a two-step process. First, student-level enrollment data need to be summarized by assigning course sections into content areas. (This may require developing a case logic that identifies the content area of every course number.) Second, counting the number of students per teacher grouped by grade and content area will reveal when teachers have students across grades and content areas. Analysis of the middle school enrollment data revealed that:

- 20% of middle school math and science teachers taught students within a single grade and single content area;
- 60% taught students across grades, but not across content areas;
- 10% taught students within a single grade, but in both math and science courses; and
- 10% taught across grades *and* across content areas.

If this district implemented a performance-pay system that rewarded individual teachers for work in core content areas, 80% of middle school teachers would be teaching more than one grade or content area and could be eligible for more than one award. This suggests that districts should carefully decide how they will determine awards for teachers of multiple subjects and grades and clearly communicate the eligibility rules to teachers.

Potential Solutions to Data Quality Challenge #3

Solutions in this example are a little less clear-cut because school systems often have to limit the number of awards that teachers can receive. One solution is to give a teacher an award for either 6th grade math or 7th grade math, but not both. Perhaps the easiest solution is to give a teacher a performance award for any grade or content area in which student performance meets specified criteria. Before doing so, however, the district should determine how this policy would affect the number of potential awards and program costs.

Another option would be to assign a teacher to the grade and content area in which he/she taught the most students. However, this might raise concerns about the compensation system, especially if teachers are assigned to high-need areas outside of their area of specialty. Although many human resource systems will note an area of expertise for teachers (e.g., secondary math), these data are often at odds with teacher assignment data in student information system. We strongly caution districts that these data should not be treated as reliable until their accuracy has been verified. Regardless of the solution that is adopted, we encourage districts to determine the number of teachers that teach multiple grades and content areas in order to project maximum costs of performance-based compensation systems accurately.

Summary

It is clear that each school system that wishes to put a pay-for-performance system in place has unique IT needs and varying capacity to retool systems to support such work. However, all school systems require high-quality data to implement an effective performance-based compensation system. Improving data quality requires understanding both social and technical roots, and efforts for improvement may be short- and long-term. The six data quality dimensions presented in this module can help school systems assess and improve their data quality, and we recommend that they begin having conversations about these dimensions with IT staff. Often, only those who work closely with an information system will know enough detail about how data are collected, stored, and organized to provide an accurate assessment of data quality challenges and effective solutions.

End Notes

[1] Watson, J.G. (2007). *Defining data quality for decision support systems in education*. Published in the proceedings of the ISMIS National and International Workshop on School Information Systems and Data Based Decision Making.

[2] English, L. (2002). The essentials of information quality management. *DM Review, 12*(9), 34-44.

[3] Thorn, C.A. (2001). Knowledge management for educational information systems: What is the state of the field? *Education Policy Analysis Archives, 9*(47). http://epaa.asu.edu/epaa/v9n47/

[4] English, L. (2002). The essentials of information quality management. *DM Review, 12*(9), 34-44.

[5] Guilford County Schools. (2007, October). Student-linkage verification. *Mission Possible Newsletter, 1*(2), 2. http://www.gcsnc.com/depts/mission_possible/pdf/October%202007%20Newsletter.pdf